UNIVERSITY
OF MANNHEIM
School of Business Informatics
and Mathematics

Chair of Practical Computer Science IV:
Dependable Systems Engineering

Master's Thesis

Vulnerabilities in Privacy-Preserving Record Linkage: The Threat of Dataset Extension Attacks

as part of the degree program Master of Science Business Informatics submitted by

Marcel Mildenberger
Matriculation number 1979905

on February 11, 2025.

Supervisor:	Prof. Dr. Frederik Armknecht PhD Student Jochen Schäfer

Abstract

The abstract should serve as an independent piece of information on your Thesis conveying a concise description of the main aspects and most important results. It should not be excessively long.

Write the abstract.

Contents

Abstract iii

1. Introduction 1
   - 1.1. Motivation 4
   - 1.2. Related Work 5
   - 1.3. Contribution 5
   - 1.4. Organization of this Thesis 6

2. Background 7
   - 2.1. Overview of PPRL 7
   - 2.2. Key encoding techniques 7
     - 2.2.1. Bloom Filters 7
     - 2.2.2. Tabulation MinHash 7
     - 2.2.3. Two-Step Hashing 7
   - 2.3. Graph Matching Attacks 7

3. Methodology 9
   - 3.1. Conceptual framework for Dataset Extension 9
   - 3.2. Implementation 9

4. Results 11
   - 4.1. Analysis 11
   - 4.2. Discussion 11

5. Conclusion 13
   - 5.1. Summary 13
   - 5.2. Future Work 13

Bibliography 15

A. Auxiliary Information 17

Eidesstattliche Erklärung 19

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

Acronyms

AI Artifical Intelligence

BF Bloom Filter

DEA Dataset Extension Attack

GMA Graph Matching Attack

ML Machine Learning

PII Personally Identifiable Information

PPRL Privacy-Preserving Record Linkage

TMH Tabulation MinHash Encoding

TSH Two-Step Hash Encoding

1. Introduction

Data and record linkage is an important component for research, software development and software projects. The primary reason for integrating data from diverse sources is to generate richer, more comprehensive insights about the same entity. Initially, deterministic record linkage, which relies on exact matches across predefined identifiers like unique id's, was the mainly used method in early linkage procedures. However, deterministic approaches often fail in real-world scenarios where data may suffer from inconsistent formatting, typographical errors, or missing values, making exact matches impossible [HSW07].

The introduction of a probabilistic record linkage framework by Fellegi et al. 1969 [FS69] marked a significant advancement in overcoming these limitations. Their work, "A Theory for Record Linkage" [FS69], proposed a statistical method for linking records across datasets by calculating the probability that two records refer to the same entity, even when inconsistencies are present in the data. Their approach evaluates common attributes assigning weights based on the likelihood of a match versus a coincidental similarity. By accounting for the discrepancies in real-world data, the so called Fellegi-Sunter model became a foundational methodology for data linkage, particularly in heterogeneous and distributed data environments where traditional deterministic methods fall short.

Such a probabilistic record linkage approach is important in sectors such as healthcare and the social sciences, where data is often distributed across multiple institutions and sources and no unique identifiers exist. In these fields, the ability to integrate datasets is essential for gaining insights and improving outcomes. For example, in the United States, the healthcare system is highly fragmented, consisting of numerous independent entities such as hospitals, clinics, insurance providers, public health agencies, and research institutions. Each of these organizations collects and stores patient data independently, often using different systems and especially formats. This fragmentation creates significant challenges when attempting to track patient outcomes, monitor disease outbreaks, or evaluate the effectiveness of treatments across the entire population. Effective data linkage can bridge these gaps by connecting records that refer to the same individual across multiple datasets. This integration is critical for tasks such as epidemiological research, public health surveillance, personalized medicine, and healthcare quality improvement [PSZ+24; VSCR17].

For instance, during the COVID-19 pandemic, the inability to efficiently link data between testing centers, hospitals, and vaccination sites hindered the timely tracking of infection rates and vaccination outcomes. Had more robust data linkage mechanisms been in place, public health officials could have responded more effectively to outbreaks and tailored interventions to specific populations. Thus, record linkage plays a critical role in transforming fragmented data landscapes into unified and actionable insights, leading to more informed decision-making. In response to the COVID-19 pandemic, organisations such as the Centers for Disease Control and Prevention and the Food and Drug Administration have launched projects to address these challenges and further develop linkage techniques [PSZ+24].

In scenarios such as the COVID-19 pandemic, data integration efforts often involve linking records on natural persons from multiple sources. For example, integrating data from different healthcare providers, laboratories and public health agencies typically requires the use of pseudo-identifiers derived from Personally Identifiable Information (PII) such as names, dates of birth or other sensitive information. However, reliance on PII for linkage raises significant privacy concerns, as improper handling of such data can lead to re-identification of individuals, with potentially serious consequences such as data breaches, identity theft or unauthorised access to personal health information [PSZ+24; SBR09].

The increasing digitization of personal data has already led to large-scale privacy breaches, demonstrating the risks of improperly secured data. Notable incidents, such as the Cambridge Analytica scandal, where personal data was misused for political profiling, highlight the ethical and regulatory challenges in data integration [IH18]. Similarly, healthcare data leaks have raised concerns about the implications of unauthorized access to medical histories, genetic data, and insurance records. Leaks of personal healthcare data can lead to severe consequences, including blackmail, discrimination, and scams, which can cause significant personal harm. For instance, individuals whose medical histories are exposed may face discrimination in employment or insurance, while others may become targets for scams exploiting their health conditions. The potential for such misuse underscores the critical importance of robust data protection measures by working with PII [Smi16].

To address these privacy risks, various techniques have been developed to protect PII during the linking process, primarily by encrypting the data prior to linking. However, the use of encrypted PII as pseudo-identifiers presents additional challenges. The key question is how to efficiently encrypt sensitive information while maintaining the ability to accurately match records [SBR09].

Therefore, Privacy-Preserving Record Linkage (PPRL) techniques are designed to facilitate data integration without exposing sensitive information, ensuring that datasets can be securely linked across different entities. In order to still be able to perform linkage while preserving privacy, similarity preserving encodings are applied to the PII. Without such similarity preserving encryption, matches between encrypted entities in different databases would not be possible [SBR09; VSCR17].

Over time, three main privacy-preserving encoding schemes have emerged as enablers for PPRL [SAH24; VCRS20].

Bloom Filter (BF) encoding is the most widely used technique in PPRL and is often considered the reference standard [SAH24]. Originally introduced by Burton Bloom in 1970 as a probabilistic data structure for efficient set membership testing [Blo70], BFs were later adapted for PPRL due to their simplicity and efficiency in both storing and computing set similarities. Their compact representation and probabilistic nature make them ideal for scalable PPRL systems, particularly in environments dealing with large datasets [SBR09]. The seminal work by Schnell et al. demonstrated the application of BFs in PPRL, particularly within healthcare settings, highlighting their ability to perform secure record matching without exposing sensitive identifiers [SBR09]. However, BFs are not without limitations. Their vulnerability to graph-based attacks and pattern exploitation has driven research into enhancing their security. Techniques such as diffusion have been proposed to obscure recognizable patterns and increase security [AHS23; SAH24]. For instance, Armknecht et al. [AHS23] explored methods to reinforce the security of BF by appending a linear diffusion layer to the BF based PPRL approach, which complicates pattern mining attacks.

To address some of these vulnerabilities to BFs, Tabulation MinHash Encoding (TMH) has been introduced as a more secure alternative. MinHash, first developed by Broder 1997 for estimating set similarities in large-scale document collections [Bro97], was adapted using tabulation-based hashing [Smi17]. Although less commonly used than BFs, TMH offers distinct advantages, including stronger security guarantees against re-identification attacks. However, these benefits come at the cost of increased computational complexity and greater memory usage, which may limit its applicability in resource-constrained environments [Smi17].

A further advancement in encoding techniques is the introduction of Two-Step Hash Encoding (TSH), which aims to combine the strengths of both BFs and TMH while mitigating their respective weaknesses. As detailed by [RCS20], TSH employs a two-phase process: data is first encoded using multiple BFs, followed by an additional hashing layer that transforms the encoded data into a set of integers suitable for similarity comparisons. This layered approach enhances privacy by adding an extra layer of obfuscation, making it more resistant to attacks while maintaining efficient similarity computations [RCS20; VCRS20].

In practice, BF-based PPRL has become the dominant standard and is widely used in areas such as crime detection, fraud prevention, and national security due to its balance of efficiency and ease of implementation. However, BF-based PPRL systems are not without limitations and vulnerabilities. Previous research has shown that there are several attacks targeting PPRL systems, with a focus on exploiting the weaknesses inherent in BF encodings. These attacks specifically target vulnerabilities in BF constructions, such as the weaknesses introduced by double hashing, structural flaws in the filter design, and susceptibility to frequent pattern-mining techniques. Notably, no specific attacks have been developed for TMH or TSH encodings, suggesting that research has focused primarily on the more widely used BF scheme. [VCRS20]

However, a more recent and practical attack has emerged that exploits vulnerabilities common to all PPRL encoding schemes. The Graph Matching Attack (GMA) uses publicly available data, such as telephone books, to re-identify encrypted individuals based on overlapping records between plaintext and encrypted databases [SAH24; VCRS20]. Unlike previous attacks that focused solely on the encoding scheme of BFs, the GMA works independently of the chosen encoding scheme. It therefore exploits the graph structure of encoded datasets to re-identify records. Given two datasets — a plaintext reference dataset and an encoded dataset — an attacker can construct similarity graphs where nodes represent individuals and edges represent similarity scores. Solving a graph isomorphism problem, attackers can infer one-to-one mappings between encoded and plaintext records, effectively breaking the privacy guarantees of PPRL. The effectiveness of GMAs depends on the overlap between the two datasets; the larger the intersection, the higher the probability of successful re-identification. While GMAs can successfully re-identify individuals present in both the plaintext and encrypted datasets, their effectiveness is limited to the overlapping subset of the two databases [SAH24; VCRS20].

This work aims to go beyond traditional GMAs by re-identifying not only individuals present in the overlapping datasets, but as many individuals as possible from the encrypted PPRL data. To achieve this, the newly introduced Dataset Extension Attack (DEA) builds on the foundations laid by GMAs. The DEA uses a neural network trained on the subset of previously re-identified individuals to predict and decode the remaining encrypted records. In doing so, the DEA significantly expands the scope and effectiveness of the attack, enabling broader de-anonymisation of PPRL datasets beyond the limitations of existing graph-based methods.

1.1. Motivation

The increasing use of PPRL in highly sensitive domains such as healthcare, finance, and national security necessitates research to validate existing techniques and ensure robust data protection [SBR09]. As data-driven applications continue to evolve, the complexity and volume of data being collected and linked across different sources grow rapid. While PPRL systems are designed to facilitate secure data integration without compromising privacy, the evolving cybersecurity threats and attacking techniques highlights the urgent need to reassess the resilience of these systems [VSCR17].

Privacy has always been a critical concern in data management, but its significance has been amplified in the era of Artifical Intelligence (AI) and Machine Learning (ML). These technologies increasingly rely on large-scale datasets that often contain sensitive PII, such as medical records, financial transactions, or behavioral data for training. If compromised, the exposure of such data can lead to severe privacy violations, including identity theft, financial fraud, and discrimination. The rise of data brokerage, where personal data is collected, aggregated, and sold—often without explicit user consent, further exacerbates privacy concerns. This commodification of personal data has made PII an attractive target for malicious actors, increasing the risk of unauthorized data linkages and re-identification attacks. As AI models become more refined, the demand for rich, high-quality data continues to grow, making data privacy an increasingly pressing issue [KM24; MK19].

In this context, the vulnerability of PPRL systems to emerging attack methods is particularly concerning. While PPRL techniques, such as BF are designed to obscure sensitive identifiers during the data linkage process, recent research has demonstrated that these systems are susceptible to GMAs. GMAs exploit the similarity-preserving properties of common encoding schemes to re-identify individuals by comparing patterns in encrypted datasets with those in publicly available plaintext datasets. This approach undermines the fundamental goal of PPRL: to protect sensitive data during the record linkage process. Although current GMAs are limited to re-identifying individuals who are present in both the encrypted and plaintext datasets, even partial data exposure in highly sensitive domains can have serious implications [SAH24; VCRS20].

The introduction of DEAs poses an even greater threat to the integrity of PPRL systems. Unlike GMAs, DEAs aim to extend the scope of re-identification to as many individuals as possible within the encrypted database. By leveraging neural networks trained on previously decoded data from GMAs, DEAs can predict and decode additional records, potentially leading to the complete de-anonymization of entire encrypted datasets. This represents a paradigm shift, as it challenges the viability of widely used PPRL techniques, such as BF-based encoding, which have been considered as secure.

The primary motivation for this research is to proactively investigate and demonstrate the consequences of such advanced attacks in order to prevent their realization in real-world scenarios. By exposing the potential vulnerabilities of PPRL systems, this work aims to highlight how attackers could exploit decrypted data to compromise privacy on a large scale. A successful implementation of the DEA will provide empirical evidence that state-of-the-art methods are insufficiently secure, emphasizing the urgent need for more robust privacy-preserving techniques.

Furthermore, there is a notable gap in the current research regarding the extension of attack capabilities beyond the intersection of datasets. While significant efforts have been made to address the vulnerabilities exposed by GMAs, there is a lack of comprehensive studies exploring how ML can be leveraged to generalize these attacks and compromise entire databases. This research aims to fill that gap by developing and evaluating the DEA, thereby contributing to the broader understanding of PPRL vulnerabilities.

By addressing this gap, this thesis seeks to contribute to the body of knowledge on PPRL vulnerabilities and serve as a foundation for future research aimed at fortifying these systems. The insights gained from this study will not only enable the development of more secure PPRL techniques but also influence best practices in data privacy and security.

1.2. Related Work

The study by Vidanage et al. [VCRS20] presents a significant advancement in the field of PPRL through the introduction of a new attack method known as the GMA. Their work begins with a comprehensive overview of PPRL systems and the similarity-preserving encoding techniques commonly employed, such as BFs. The GMA exploits vulnerabilities in these encoding schemes by leveraging their ability to retain partial similarity information even after encryption. By constructing similarity graphs from both encrypted and plaintext datasets, the GMA solves a graph isomorphism problem to align nodes and successfully re-identify individuals in the encrypted dataset using publicly available sources, like phonebooks. This method demonstrates the universal applicability of GMAs across various PPRL schemes, highlighting a critical weakness in systems that were previously considered robust [VCRS20].

Building on this foundation, Schäfer et al. [SAH24] revisited and extended the work of Vidanage et al. Their contribution lies in a meticulous reproduction and replication of the original GMA, during which they identified a critical flaw: an undocumented pre-processing step in the provided codebase that inadvertently increased the effectiveness of the attack. While this step was initially intended to enhance computation performance, it introduced errors to the proposed GMA. Schäfer et al. corrected this issue and further optimized the GMA, resulting in improved robustness and efficiency. Their enhanced implementation achieved higher re-identification rates compared to the original approach. This improvement not only validates the vulnerabilities highlighted by the GMA but also underscores the potential for refining attack methodologies to expose even greater weaknesses in PPRL systems.

The work of Schäfer et al. is particularly relevant to this thesis, as their improved GMA implementation and accompanying codebase provide the foundation for the DEA proposed in this study. While the GMA is limited to re-identifying individuals present in both encrypted and plaintext datasets, the DEA seeks to extend the scope of re-identification beyond this intersection. By leveraging neural networks trained on the re-identified individuals from the GMA, the DEA aims to predict and decode additional records, potentially leading to the complete de-anonymization of encrypted datasets.

To date, no existing research has proposed an approach comparable to the DEA. This thesis addresses this gap by developing and evaluating the DEA, thereby contributing to the broader understanding of PPRL vulnerabilities and highlighting the urgent need for more secure data linkage techniques.

1.3. Contribution

The contribution of this thesis is divided into three main parts. First, a comprehensive analysis of PPRL systems is carried out, with particular emphasis on the three major encoding schemes: BF encoding, TSH encoding, and TMH encoding. This analysis aims to highlight the basic principles, strengths and weaknesses of each encryption scheme, setting the stage for the subsequent investigation of their susceptibility to the DEA.

Next, the current state of the art GMA is analysed and its limitations are discussed in detail. Although GMAs have proven effective in re-identifying individuals within overlapping datasets, their applicability is limited to the intersection of plaintext and encrypted records. This inherent limitation highlights the need for more advanced attack strategies that can go beyond this.

The main focus of this thesis is the implementation and evaluation of the DEA, which attempts to outperform the capabilities of GMAs by decrypting a larger fraction of encrypted records. To achieve this, the thesis examines the conceptual foundations, theoretical underpinnings and technical requirements of the DEA. Building on the initial re-identifications made by the GMA, the DEA employs a supervised machine learning approach, specifically using neural networks trained on previously decoded data to predict and re-identify remaining encrypted records. This method significantly extends the scope of de-anonymisation in PPRL systems and provides a novel approach to current research.

The DEA is then evaluated against the three major PPRL encoding schemes. While the specific encoding method has minimal impact on the GMA, which relies primarily on solving a graph isomorphism problems, it plays a role in the DEA. This is due to the fact that the neural network must be trained separately for each encoding scheme to account for the unique structural characteristics and encoding nuances. However, the DEA is designed with adaptability in mind, ensuring that it can be effectively applied across different encoding schemes, thus increasing its generalisability and practical relevance.

Through this research, the thesis aims to answer critical questions about the robustness of PPRL systems. It investigates how effective supervised machine learning-based DEAs are at re-identifying the remaining entries that GMAs cannot decode. It also examines how different encoding schemes affect the performance and accuracy of the DEA, providing insight into which schemes are more susceptible to such attacks and why. By addressing these questions, the thesis contributes to a deeper understanding of the vulnerabilities inherent in PPRL systems and lays the groundwork for the development of more secure privacy-preserving techniques.

1.4. Organization of this Thesis

This thesis is divided into four main sections: technical background, methodology, results, and conclusion.

First, an overview of PPRL systems is given, with particular emphasis on a thorough analysis of the most commonly used encoding techniques. Next, the existing GMA is introduced and explained in order to provide the basis for the study. In addition, an overview of neural networks is given to provide the necessary background knowledge.

Next, a detailed description of the attack model for the DEA is outlined, including how neural networks are used to enhance the attack. This is followed by an explanation of the actual implementation of the DEA, along with a discussion of the experiments conducted. The results of the DEA on different encryption schemes are then analysed and evaluated. Finally, the thesis concludes with a summary of the main contributions, a discussion of the broader implications, and suggestions for future research.

2. Background

2.1. Overview of PPRL

2.2. Key encoding techniques

2.2.1. Bloom Filters

2.2.2. Tabulation MinHash

2.2.3. Two-Step Hashing

2.3. Graph Matching Attacks

3. Methodology

3.1. Conceptual framework for Dataset Extension

3.2. Implementation

4. Results

4.1. Analysis

4.2. Discussion

5. Conclusion

5.1. Summary

5.2. Future Work

Bibliography

[AHS23] Frederik Armknecht, Youzhe Heng, and Rainer Schnell. "Strengthening privacy-preserving record linkage using diffusion." In: *Proceedings on Privacy Enhancing Technologies* (2023).

[Blo70] Burton H Bloom. "Space/time trade-offs in hash coding with allowable errors." In: *Communications of the ACM* 13.7 (1970), pp. 422–426.

[Bro97] Andrei Z Broder. "On the resemblance and containment of documents." In: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE. 1997, pp. 21–29.

[FS69] Ivan P Fellegi and Alan B Sunter. "A theory for record linkage." In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1183–1210.

[HSW07] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Vol. 1. Springer, 2007.

[IH18] Jim Isaak and Mina J Hanna. "User data privacy: Facebook, Cambridge Analytica, and privacy protection." In: *Computer* 51.8 (2018), pp. 56–59.

[KM24] Jennifer King and Caroline Meinhardt. *Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World*. White Paper, Stanford University Institute for Human-Centered Artificial Intelligence (HAI). 2024. URL: http://www.darkpatternstipline.org.

[MK19] Karl Manheim and Lyric Kaplan. "Artificial intelligence: Risks to privacy and democracy." In: *Yale JL & Tech.* 21 (2019), p. 106.

[PSZ+24] Aditi Pathak, Laina Serrer, Daniela Zapata, Raymond King, Lisa B Mirel, Thomas Sukalac, Arunkumar Srinivasan, Patrick Baier, Meera Bhalla, Corinne David-Ferdon, et al. "Privacy preserving record linkage for public health action: opportunities and challenges." In: *Journal of the American Medical Informatics Association* 31.11 (2024), pp. 2605–2612.

[RCS20] Thilina Ranbaduge, Peter Christen, and Rainer Schnell. "Secure and accurate two-step hash encoding for privacy-preserving record linkage." In: *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*. Springer. 2020, pp. 139–151.

[SAH24] Jochen Schäfer, Frederik Armknecht, and Youzhe Heng. "R+R: Revisiting Graph Matching Attacks on Privacy-Preserving Record Linkage." In: *Proceedings of [Conference Name, if available]*. Available at: https://github.com/SchaeferJ/graphMatching. University of Mannheim. 2024.

[SBR09] Rainer Schnell, Tobias Bachteler, and Jörg Reiher. "Privacy-preserving record linkage using Bloom filters." In: *BMC medical informatics and decision making* 9 (2009), pp. 1–11.

[Smi16] Tanshanika T Smith. *Examining data privacy breaches in healthcare*. Walden University, 2016.

[Smi17] Duncan Smith. “Secure pseudonymisation for privacy-preserving probabilistic record linkage.” In: *Journal of Information Security and Applications* 34 (2017), pp. 271–279.

[VCRS20] Anushka Vidanage, Peter Christen, Thilina Ranbaduge, and Rainer Schnell. “A graph matching attack on privacy-preserving record linkage.” In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 1485–1494.

[VSCR17] Dinusha Vatsalan, Ziad Sehili, Peter Christen, and Erhard Rahm. “Privacy-preserving record linkage for big data: Current approaches and research challenges.” In: *Handbook of big data technologies* (2017), pp. 851–895.

A. Auxiliary Information

Eidesstattliche Erklärung

Hiermit versichere ich, dass diese Abschlussarbeit von mir persönlich verfasst ist und dass ich keinerlei fremde Hilfe in Anspruch genommen habe. Ebenso versichere ich, dass diese Arbeit oder Teile daraus weder von mir selbst noch von anderen als Leistungsnachweise andernorts eingereicht wurden. Wörtliche oder sinngemäße Übernahmen aus anderen Schriften und Veröffentlichungen in gedruckter oder elektronischer Form sind gekennzeichnet. Sämtliche Sekundärliteratur und sonstige Quellen sind nachgewiesen und in der Bibliographie aufgeführt. Das Gleiche gilt für graphische Darstellungen und Bilder sowie für alle Internet-Quellen.
Ich bin ferner damit einverstanden, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs in elektronischer Form anonymisiert versendet und gespeichert werden kann.

Datum

Marcel Mildenberger